Speech rate calculations with short utterances

Speech Rate Calculations with Short Utterances: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task

Hayakawa Akira, [1] **Carl Vogel,** [1] **Saturnino Luz,** [2] **Nick Campbell** [1]
[1] School of Computer Science and Statistics, Trinity College Dublin, Ireland
[2] Usher Institute of Population Health Sciences & Informatics, University of Edinburgh, UK
campbeak@tcd.ie, vogel@cs.tcd.ie, S.Luz@ed.ac.uk, nick@tcd.ie

Abstract

The motivation for this paper is to present a way to verify if an utterance within a corpus is pronounced at a fast or slow pace. An alternative method to the well-known Word-Per-Minute (wpm) method for cases where this approach is not applicable. For long segmentations, such as the full introduction section of a speech or presentation, the measurement of wpm is a viable option. For short comparisons of the same single word or multiple syllables, Syllables-Per-Second (sps) is also a viable option. However, when there are multiple short utterances that are frequent in task oriented dialogues or natural free flowing conversation, such as those of the direct Human-to-Human dialogues of the HCRC Map Task corpus or the computer mediated inter-lingual dialogues of the ILMT-s2s corpus, it becomes difficult to obtain a meaningful value for the utterance speech rate. In this paper we explain the method used to provide a alternative speech rate value to the utterance of the ILMT-s2s corpus and the HCRC Map Task corpus.

Keywords: speech rate, utterance duration comparison, task oriented dialogues

1. Introduction

Computer mediated communication is becoming more frequent. The next step in this new communication style, is Inter-Lingual Computer Mediated Communication. Recently, Microsoft released the Skype Translator (Lewis, 2015) that translates up to 10 languages in Speech-to-Speech Machine Translation. The Japanese Ministry of Internal Affairs and Communication has announced that the Tokyo Olympics in 2020 are to employ information systems that use multilingual machine mediated communication in Speech-to-Speech (S2S) form for 17 languages, using the VoiceTra system (Matsuda et al., 2013). Computer mediated multi-lingual communication will create a conversation style that few users have adequate exposure (Hara and Iqbal, 2015) and may have difficulty adapting to. For example, using the speech rate calculation method explained in this paper, gender differences in the adaptation speech rate have been identified (Hayakawa et al., 2017b) but with little to no improvement to the automatic speech recognition results. With users of computer mediated interaction, the "Speakers are unlikely to have a good model of what computers are likely to know, presumably in part because natural language systems tend to be tailored to particular and quite specific uses." (Branigan et al., 2010, p. 2366). Being able to provide feedback to the user about the changes to their speech rate, should help prevent the hyper-articulation reported by Hayakawa et al. (2017b).

The method in this paper is not new in theory. Sztahó et al. (2015) use a similar method of comparing the syllable duration it takes a subject to utter a specific phrase with previous recordings of the utterance by the same subject. The method explained in this paper has also been used to calculate the speech rate in previous publications by the author (Hayakawa et al., 2015; Hayakawa et al., 2016a; Hayakawa et al., 2017a), but continues to attract interest and has yet to be explained in detail.

2. The Dataset

Two corpora have been used in this analysis, the ILMT-s2s corpus (Hayakawa et al., 2016b) and the HCRC Edinburgh Map Task corpus (Anderson et al., 1991).

The ILMT-s2s corpus: The corpus consists of 15 dialogues of English speakers communicating with Portuguese speakers to perform the HCRC Edinburgh Map Task (Anderson et al., 1991), a task where the subject is to guide the interlocutor along a predefined route on the map of one of the subjects. The subjects are situated in different rooms and communicate in their mother tongue to their interlocutor using a Speech-to-Speech Machine Translation (S2S-MT) system that Hayakawa et al. (2016b) call the *ILMT-s2s System*. The corpus consists of ≈9.5 hours of audio, video and biological signal recordings of interlingual system mediated communication of 15 subject pairs (15 English and 15 Portuguese speakers).

The HCRC Edinburgh Map Task corpus: The corpus consists of 128 English dialogues of direct human-to-human task based interactions using the map task technique to elicit the comunication. The recordings were split in two settings, with half the subjects being able to see their interlocutor's face (i.e., with eye-contact), while the other half had screens placed between them (i.e., without eye-contact). To standardise the data, only dialogues that used the same maps (maps 1 & 7) as those used in the ILMT-s2s corpus were kept for this study, resulting in a total of only 16 dialogues out of the 128 (half male, half female).

3. Calculating the Speech Rate

To calculate if any given utterance in the ILMT-s2s corpus was uttered quickly or slowly, the duration of the subject utterance was compared with that of the same utterance repeated by the TTS system used in the ILMT-s2s System during the collection of the corpus.

This method was chosen for two reasons. The first, was because we wanted to verify if the subjects were aligning their speech rate to the ILMT-s2s System output. This would

result in the subject utterance aligning with the 180 wpm speech rate setting of the ILMT-s2s System's TTS output. The second reason was because the utterances in the ILMT-s2s corpus were expected to be low in word count and short in duration from the pre-corpus collection test runs, and hence, make it difficult to get a meaningfull speech rate value. This expectation was materialised in the ILMT-s2s corpus collection with the median word count per utterances of 4 words and a median duration per utterances of 1.50 seconds as indicated in Table 1.

Word Count	*Mdn*	*M*	*SD*	Count
ILMT-s2s All Subjects	4	5.168	6.01	3,628
ILMT-s2s English Subjects	4	4.919	6.76	1,980
ILMT-s2s Portuguese Subjects	4	5.466	4.97	1,648
Duration (sec.)	*Mdn*	*M*	*SD*	Count
ILMT-s2s All Subjects	1.493	2.244	2.75	3,628
ILMT-s2s English Subjects	1.285	1.939	2.96	1,980
ILMT-s2s Portuguese Subjects	1.874	2.610	2.49	1,648

Table 1: Summary of word count and duration in corpus

A summary of the rate that each utterance in the ILMT-s2s corpus was uttered is reported in Table 2 using the measure of wpm. The utterances are indicated to have been

Speech Rate (wpm)	Min.	*Mdn*	*M*	Max.	*SD*
ILMT-s2s All Subjects	15.24	141.00	153.40	618.60	65.28
ILMT-s2s En. Subjects	16.61	155.30	167.70	618.60	68.75
ILMT-s2s Pt. Subjects	15.24	127.90	136.20	459.60	56.26

Table 2: Summary of wpm per utterance

spoken as slowly as 15.40 wpm and as fast as 618.60 wpm with an arithmetic mean speed of 153.4 wpm and a *SD* of 65.28. Of this range, 60% of the utterances were reported to have been uttered within a 100 wpm scope between 100.03 wpm (20^{th} percentile) and 200.00 wpm (80^{th} percentile). Though these wpm values indicate that the subject uttered their speech relatively slowly, it is not clear yet if the sample utterance was intentionally spoken slowly or fast, or if the wpm calculation method is providing numbers that only represent slow or fast speech rates.

If the orthodox way of calculating the Words Per Minute (wpm) as $(W/T \times 60)$, where W is the word count per utterance and T is the duration of the utterance in seconds was used, the imbalanced nature of the spoken words would create a variance that is presumably difficult to interpret. One example of this is the different wpm value of the three single word utterances, "Perfect", "Yes" and "Okay" that are indicated in Table 3. The maximum wpm value for the

Speech Rate (wpm)	Min.	1st	*Mdn*	*M*	3rd	Max.
Perfect	98.2	104.0	110.5	113.4	123.7	130.4
Yes	73.1	117.9	138.9	135.8	153.1	183.5
Okay	55.5	112.2	139.9	154.0	173.9	618.6
Pebbled shore	98.2	100.1	101.7	106.9	111.2	120.7
Go down	98.4	121.6	129.6	124.5	133.1	136.1
Then where?	206.9	213.1	219.4	216.7	221.6	223.9

Table 3: Summary of single and two word subject utterance duration and wpm sample

word "Perfect" is 130.40 wpm, but did the subject utter this word at the same intentional rate as the median wpm value for "Yes" at 138.90 wpm or "Okay" at 139.90 wpm? Unlike measurements for traveling distance speeds of kilometres per hour (kph) or miles per hour (mph) where the 2 items of distance and time are consistent, the length of distance that is measured as kilometres or miles, and the duration of time that is measured as hours, the measurement of wpm only has 1 consistent item, the durations that is measured in minutes. The measurement of a word is not of a consistent length — one might say that this variable length in words is a characteristic of the "stress-timed" English and Portuguese languages, but the variability in duration has also been found in a theoretically isochronous "syllable-timed" language such as Japanese, too (Arai and Greenberg, 1997). Since the length of the single word is different, it is not possible to know if the 130.40 wpm value for the single word utterance of "Perfect" is fast or slow when compared with the given example of "Yes" or "Okay", as it is possible with kph/mph where the numerator as well as the denominator are of a consistent measurement. This is not just a problem of single word utterances, but also in multiple word utterances as indicated in Table 3 with the different wpm values of the three multi word utterances. The median wpm values for "Pebbled shore" (an item on the map), "Go down" and "Then where?" are 101.70 wpm, 129.60 wpm and 219.40 wpm respectively. However there is no way of knowing if the difference in the wpm values of these three multi word utterances and even the single word utterances in Table 3 are the result of a speech rate difference or the difference in the duration to pronounce the given word. The bar chart

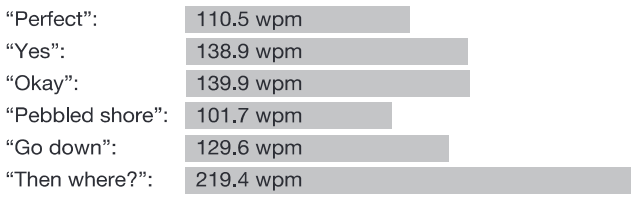

Figure 1: *Mdn* wpm for the six samples — Subject

in Figure 1 provides no more information than five words were spoken slowly and one was not.

The concept of wpm as a reference of speech rate is already an average of word combinations within the utterance of a minute. The random mixture of words with long and short duration are mixed into an utterance that is then calculated into a quantitive measure. Though the frequency of word duration has been previously investigated in written language (Zipf, 1945) and also in spoken language (Greenberg, 1999; Batliner et al., 2001; Bell et al., 2002; Bell et al., 2009), given the varying short sentences, in the unscripted speech of the ILMT-s2s corpus, the irregular duration of a word used in the short phrases would not provide an accurate wpm value to provide a reliable quantitive measure. If the window were a longer period of time, such as the first quarter of the dialogue, there might be enough words to balance out the variance, but not in this situation where the window is a single utterance. The other method of using *syllables per second* as a measure may be more reliable than wpm, however this would require highly trained

annotators to phonologically segment the data. With this comes a new issue of whether the utterance is annotated as indicated in the dictionary or following the actual phonetic sound that was uttered.[1] A fully unsupervised method of calculating the speech rate estimation based on syllable nuclei detection was presented by de Jong and Wempe (2009) as a Praat (Boersma and van Heuven, 2001) script, though we have yet to test this method on the data of the ILMT-s2s corpus to verify whether similar reliability can be obtained. However, analysis of the *syllables per second* may not provide any better results since an analysis of the Switchboard corpus data (Godfrey et al., 1992), which is a corpus of a collection of natural and spontaneous telephone conversations, by Greenberg (1999, p. 167) indicated the following:

> Although only 22% of the Switchboard lexicon is composed of monosyllabic forms, approximately 80% of the corpus tokens are just one syllable in length [...]. The portion of the lexicon consisting of three or more syllables (38%) is rarely exhibited in spontaneous language, accounting for less than 5% of the spoken instances [...].

This would indicate that even using the measure of syllables per second, the length of the dialogues in the ILMT-s2s corpus may be too short and the problem identified with the wpm measure may also continue to persist.

Since the main objectives of the studies with the ILMT-s2s corpus were to investigate the alignment or adaptation of the subject to the ILMT-s2s System, it was perceived that a comparison with the ILMT-s2s System's TTS speech rate would be adequate as it would provide a one-to-one reference point with the utterances of the ILMT-s2s corpus. The rationale behind this idea was to recreate every utterance that was uttered by the subjects of the ILMT-s2s corpus using the same TTS output system used in the ILMT-s2s System and compare like-for-like utterance durations between the human subject and the ILMT-s2s System's TTS voice. This method would theoretically remove the variance in the resulting speech rate value that is created by the differing duration in pronouncing words of differing lengths, since the reference will also be pronouncing the same word — therefore, creating a more stable speech rate. To compare the speech rate, the transcription text of all speakers was output to a plain text file that was then read out by the TTS system at a speech rate setting of 180 wpm and saved as an audio file. Using Praat, the audio file was segmented manually by the first author using the transcription from the plain text file to provide the boundaries of TTS output of each utterance. Once completed, the duration of each TTS output reference was calculated from the start and end times of the segmented audio files. A summary of the duration and wpm of each TTS output is as reported in Table 4.

4. Analysis

As with the subject utterances wpm values in Table 1, the TTS output data show a big variance between the slow output measures and the fast output measures, but with a smaller *SD* when compared to the subject utterance measures. The TTS output are as slow as 45.09 wpm and as fast as 536.10 wpm with an arithmetic mean speed of 185.50 wpm and a *SD* of 56.79. Of this range, 60% of the TTS output were reported to have been within a ≈100 wpm scope between 137.57 wpm (20th percentile) and 231.69 wpm (80th percentile) which is similar to the subject utterance range reported earlier. This summary

Duration (sec.)	Min.	*Mdn*	*M*	Max.	*SD*
ILMT-s2s All Subjects	0.112	1.250	1.694	56.500	1.90
ILMT-s2s En. Subjects	0.166	1.124	1.508	56.500	2.03
ILMT-s2s Pt. Subjects	0.112	1.491	1.917	14.240	1.69
Speech Rate (wpm)	**Min.**	***Mdn***	***M***	**Max.**	***SD***
ILMT-s2s All Subjects	45.09	179.40	185.50	536.10	56.79
ILMT-s2s En. Subjects	45.82	190.90	192.90	390.20	58.50
ILMT-s2s Pt. Subjects	45.09	170.10	176.60	536.10	53.33

Table 4: Summary of duration and wpm of the subject utterances output by the TTS system

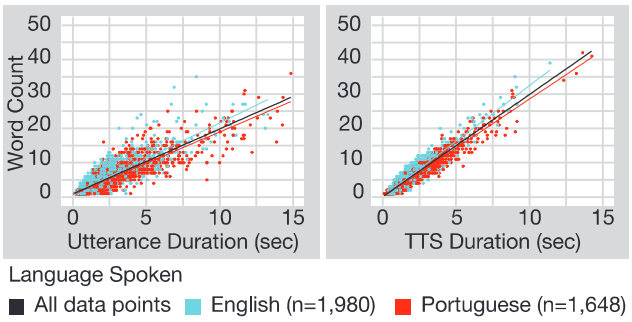

Figure 2: Scatter plot of ILMT-s2s corpus utterance durations by word count

in itself does not show much, but a simple linear regression model of the two wpm values highlights and confirms the differences between the two speech rates (Figure 2). The regression equation of the word count and the duration for the subject utterance is "$Word\ Count = 0.61648 + 2.02693\ Utterance\ duration$", with an $Adjusted\ R^2$ value of 0.858 ($p < 2.2e-16$) and for the TTS output it is "$Word\ Count = -0.06870 + 3.08959\ TTS\ output\ duration$", with an $Adjusted\ R^2$ value of 0.948 ($p < 2.2e-16$). There are two interesting details from these results that illustrate the data: (i) the slope difference of 1.062 which predicts that the speech rate of the TTS output will be ≈ 30% faster than the subject utterance, and (ii) the 0.948 $Adj\ R^2$ value that indicates there is a variance in the speech rate of the TTS output since it is not the maximum value of 1.0, but from the 0.09 difference, this variance is smaller than that of the subject utterances. These two points are retrievable from the summary of the data in Table 2, and Table 4, however the linear regression equation and Figure 2 represents this more clearly. Although this identifies the speech rate difference and variation difference of the two measurements of subject utterance speed and TTS output speed, it does not help clarify the speech rate differences of the six single and multiple word examples provided in Table 3. The main difference

[1] Greenberg (1999, p. 163) reported eighty different pronunciations of the word "and" in a study of the Switchboard Transcription Corpus (Greenberg, 1997)

that can be observed when comparing the one and two word example phrases of the subject utterances and TTS output, is again the reduction in the variability of the speech rate, the *SD* values (Table 4). This reduction indicates the regularity in which the phrase is repeatedly output by the TTS system. Which is an effect that is completely expected since this is the exact reason that the TTS output was used; to provide a standardised, regular speech rate. However, the question of which phrase is spoken fast or slow is not identifiable from this data. The bar chart in Figure 3 does not provide any better understanding of the speech rate of short utterances.

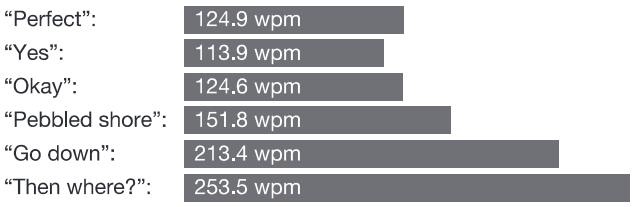

Figure 3: *Mdn* wpm for the six samples — TTS output

Taking advantage of the regular speech rate of the TTS output, the TTS output duration can be used as a reference that the subject utterance is compared against. By dividing the subject utterance duration by the TTS output duration, a value that identifies the difference from the regular speech rate obtained by the TTS output can be retrieved with an equation of (S/T), where S is the duration of the speaker's utterance and T is the duration of the TTS output. The resulting values of this measurement method are reported in Table 5. The range between zero

Speech Rate	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.
ILMT-s2s All	0.167	0.982	1.233	1.379	1.588	16.400
ILMT-s2s En	0.167	0.925	1.188	1.287	1.525	9.759
ILMT-s2s Pt	0.414	1.044	1.291	1.490	1.670	16.400

Table 5: *Summary of subject utterance duration divided by TTS output duration — "S/T"*

(0) and one (1) are values from utterances that were spoken faster than the TTS output and the range of one (1) and above are utterances that were spoken slower than the TTS output. The *Mdn* value being slower than the TTS output can also be confirmed from the slope of the regression equation of "$Subject\ utterance\ duration = 0.239752 + 0.648037\ TTS\ output\ duration$" with an $Adjusted\ R^2$ value of 0.883 ($p < 2.2e - 16$) as indicated in the graph on the right in Figure 4. With the slope being smaller than one (1), the regression equation is indicating that the subject utterance duration will not be shorter than the TTS output duration, and therefore the majority of subject utterance will be slower than the TTS output. This is evident from the lack of values within the pink coloured region of Figure 4, which covers the area of a slope greater than one (1). Unfortunately understanding the figures of Table 5 requires a moment of mental arithmetic and it is not immediately identifiable from just a glance at the results if the utterance was spoken quickly or not. To make the measurement values immediately evident, one (1) was subtracted from the

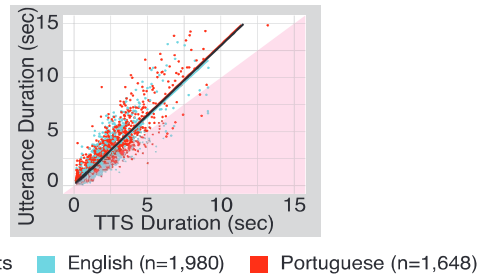

Figure 4: Scatter plot subject utterance by TTS output

resulting value to differentiate utterances that were faster or slower than the TTS output by using zero (0) as a reference division line, and the value was also multiplied by minus one (-1) to display utterances spoken slower than the TTS output as a negative value and faster utterances as a positive value. The resulting values of this conversion are summarised in Table 6 and displayed as percentage values. With this modification, we think that the values in reference to speech rate of the utterance are easier to understand as illustrated in Figure 6.

Speech Rate	Min.	1st	*Mdn*	*M*	3rd	Max.
ILMT-s2s All	−1539.0	−58.8	−23.4	−37.9	1.9	83.3
ILMT-s2s En	−878.7	−52.4	−18.8	−28.7	7.6	83.3
ILMT-s2s Pt	−1539.0	−67.0	−29.1	−49.0	−4.4	58.6

Table 6: *Summary of subject utterance duration divided by TTS output duration — "$1 - (S/T)$"*

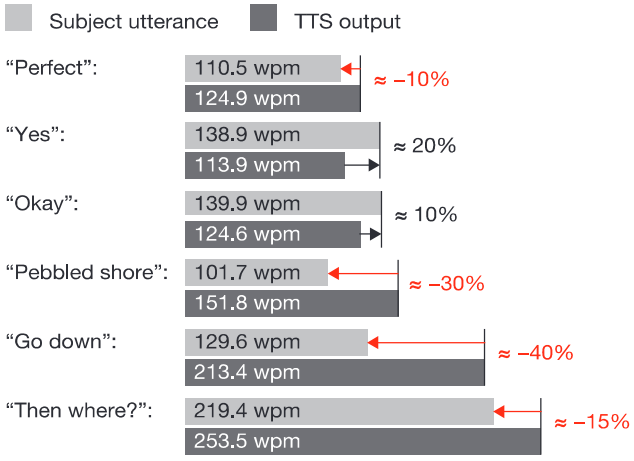

Figure 5: *Example speech rate comparison — Subject utterance & TTS output*

With this comparison, where the duration required to pronounce the word is taken into account, it is possible to say with a higher degree of certainty that the "4.22%" maximum value of the single word utterance of "Perfect" represents a speech rate that is a similar rate to the "3.45%" first quarter value of "Yes" and the "9.00%" mean value of "Okay" as listed in Table 7. Using this method of comparing the subject utterance duration and the TTS output duration, the meaning that was lacking in the wpm values in Table 3 can now be presented with a more comprehensive value — the durational difference between the subject

Difference from TTS (%)					
Utterance	Min.	1st Qu.	*Mdn*	*M*	Max.
Perfect	−27.65	−18.98	−13.61	−11.06	4.22
Yes	−55.82	3.45	18.01	12.91	38.21
Okay	−130.70	−8.14	13.54	9.00	83.30
Pebbled shore	−54.61	−52.06	−49.50	−43.29	−25.77
Go down	−117.00	−74.98	−64.39	−73.36	−56.97
Then where?	−22.55	−19.07	−15.58	−17.31	−13.81

Table 7: *Comparison with TTS duration sample — Percentage difference with TTS output duration*

utterance and a regular speech rate of the TTS output.
This method of measuring the speech rate of the subjects was repeated for the data of the HCRC Edinburgh Map Task corpus and as illustrated in Figure 6 and listed in Table 8, it is possible to see a clearer indication of the subject utterance speech rate. For example, when the measurement of wpm is used, it is not possible to say if the median (*Mdn*) utterance speed for "Okay" at 156.9 wpm was really spoken at half the intentional speed of "Got you" at 372.7 wpm. However, by comparing the utterance duration of the subject and the duration of the TTS output, both utterances can be indicated with a more reasonable median (*Mdn*) speech rate value and say that "Got you" at 42.6% was spoken with a slightly faster speech rate than "Okay" at 33.29%, but not twice as fast as the wpm value would lead one to believe.

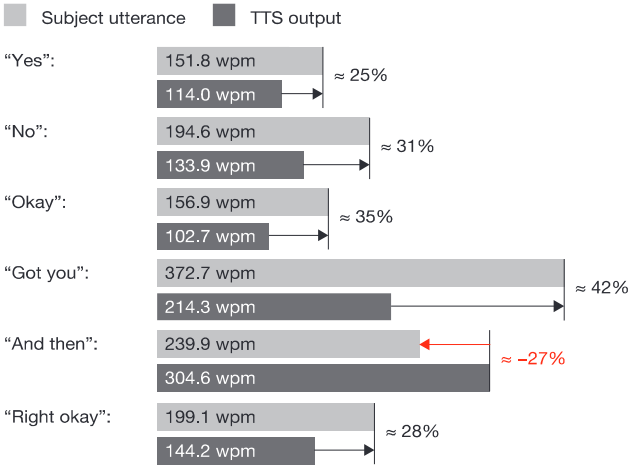

Figure 6: *Example HCRC corpus speech rate comparison — Subject utterance & TTS output*

5. Discussion and Conclusion

By comparing the speech rate of the subject utterance with a reference TTS output of the same utterance, I have indicated that the resulting value is easier to interpret and more robust than the wpm measurement method, by comparing the variable subject utterance duration with a standard measurement of the TTS output duration. Using this method, it becomes possible to visualise the speech rate of short utterances with a higher degree of certainty. For example, Figure 7 illustrates how the values from this method can illustrate how the speech rate changes from utterance to utterance within a dialogue of two different corpora, indicating how speech rates differ when talking directly to ones interlocutor (HCRC corpus) and when one's communication is

Utt. Speech Rate (wpm)						
	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.
Yes	92.7	134.2	151.8	157.3	180.7	235.6
No	80.3	164.8	194.6	238.8	287.2	653.6
Ok	77.8	134.2	156.9	169.5	189.3	421.1
Got you	308.7	351.2	372.7	381.1	395.1	485.8
And then	158.3	218.1	239.9	271.2	255.7	475.1
Right ok	87.8	165.4	199.1	214.1	243.9	481.3
Difference from TTS (%)						
	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.
Yes	−23.00	15.58	23.56	23.01	37.02	51.30
No	−66.90	18.43	31.64	32.30	52.78	79.60
Ok	−31.78	22.46	33.29	33.27	45.52	75.93
Got you	38.59	38.93	42.60	43.96	45.78	55.42
And then	−98.43	−47.24	−26.98	−23.83	−12.98	39.67
Right ok	−64.22	12.93	26.54	23.14	40.30	67.41

Table 8: *Comparison with TTS duration sample*

mediated by a Speech-to-Speech Machine Translation system.

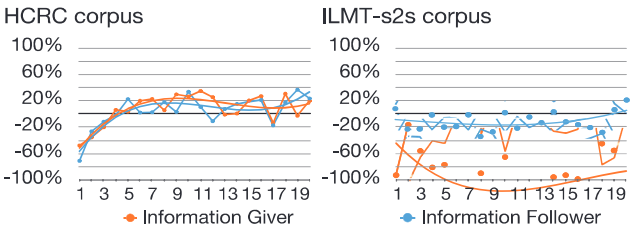

Figure 7: Example of speech rate comparison of speakers

5.1. Pause for thought

However, Yuan et al. (2006, p. 542), in their study of the Switchboard corpus (Godfrey et al., 1992) and the Chinese CallHome and CallFriend corpus (Yuan and Jurafsky, 2005) have indicated that utterances speech rates between one and seven words change drastically:

> We can see that in both English and Chinese, there is an abrupt rise of speaking rate for the segments containing from one to seven words. For the segments having eight to about 30 words, however, the speaking rate stays level. And then, especially in English, the speaking rate rises again, but with a more gradual slope.

A similar phenomenon was observed with the wpm measurements of the subject utterances (Information Giver, and Information Follower), and the TTS output of the ILMT-s2s corpus as illustrated in Figure 8. From the boxplots in Figure 8 it is possible to observe that the mean wpm speech rate increases up to utterances with a word count of 5 and then starts to fluctuate after that. Since the ILMT-s2s corpus data has indicated that there is a speech rate difference between Female (♀) and Male (♂) subjects (Hayakawa et al., 2017b), the grouping of gender was verified and illustrated as Figure 9. From Figure 9 it is possible to observe that the TTS output and Male (♂) subjects' speech rate (wpm) increase rapidly from one word to four or five words, while the Female (♀) subjects' speech rate (wpm) does not increase much and remains basically flat, which is an obser-

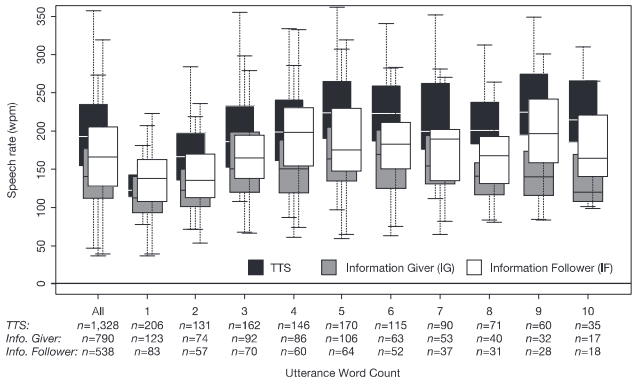

Figure 8: *Boxplot of the 3 talk types' multi word utterance speech rates in various groupings.*

vation that was not reported by Yuan et al. (2006). The

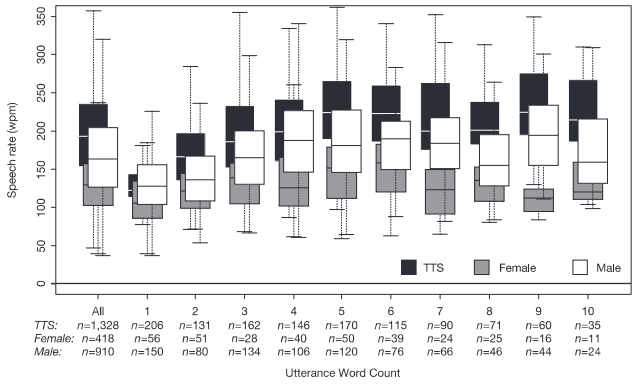

Figure 9: *Boxplot of the 3 talk types' multi word utterance speech rates in various groupings.*

gender difference reported by Yuan et al. (2006, pp. 543–544) was that the Female (♀) subjects had a slower speech rate, but with a minor difference to the Male (♂) subjects.

> Males tend to speak faster than females [. . .] difference between them is, however, very small, only about 4 to 5 words or characters per minute (2%), though it is statistically significant. It might be due to things that we would not normally think of as speech-rate parameters, such as differences in word-frequency distributions.

To verify if this pattern also exists in the HCRC Edinburgh Map Task corpus, the wpm speech rate divided by gender and role were compared as illustrated in Figure 10, with similar curves in the graphs being observed. This indicated that the data of the HCRC Edinburgh Map Task corpus also shows a curvature similar to that reported by Yuan et al. (2006) and also a curvature more similar to the TTS output than the data of the ILMT-s2s corpus. However the increase that Yuan et al. (2006) reported to start with utterances about 30 words long, can be seen to start from ≈12 words for both the subjects of HCRC Edinburgh Map Task corpus and the TTS output. It is important not to forget that the analysis by Yuan et al. (2006) was performed on the SwitchBoard corpus (Godfrey et al., 1992) which is

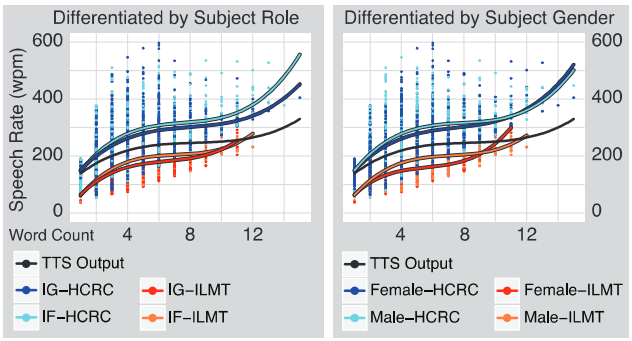

Figure 10: *Scatter plot of ILMT-s2s corpus utterance durations by word count*

a collection of telephone dialogues. Morikawa and Maesako (1998, p. 149) in their reference to the publication in Japanese of Yoshida and Kakuta of a study in audio telephone communication, indicate that the subjects of their study consider communication over the telephone as a different alternative to Face-to-Face communication with its own communication style.

> A study by Yoshida et al. showed that college students in and around Kyoto–Osaka (effective response 549) cited the following reasons for the phone's popularity: speed, available anytime, available anywhere, no visual information about the other end, and easier way to say what we want than in face-to-face conversation. They also cited the following disadvantages: no visual information about the other end and difficulty in conveying subtle emotion. How interesting that no visual information about the other end becomes both an advantage *and* disadvantage !
>
> The above study suggests that the telephone provides its own conversation environment rather than a substitute for face-to-face conversation – in short, people have found a new way to communicate using voice alone.

Due to the different communication methods and the fundamental fact that the data being analysed is different, identical results are not expected, but the pattern is present in both the HCRC Edinburgh Map Task corpus and the ILMT-s2s corpus. The results from Yuan et al. (2006) is of concern because 60% of the data in the ILMT-s2s corpus are located between a word count of 2 words (20th percentile) and 8 words (80th percentile). This range between 2 and 8 words is where Yuan et al. (2006, p. 542) mention that there is an "abrupt rise of speaking rate". The median word count for the Information Giver and the Female (♀) subjects are five and the median word count for the Information Follower and Male (♂) is four. Following this theory, the speech rate of the Information Follower should have a slower wpm speech rate than the Information Giver, and Male (♂) subjects too should have a slower wpm speech rate than the Female (♀) subjects. From Table 9 the Information Giver and Female (♀) subjects and the combination of four word Information Giver and five word Information Follower utterances follow this trend of a faster speech rate

ILMT-s2s corpus (English wpm speech rate)	Min.	1st Qu.	Mdn.	Mean	3rd Qu.	Max.	SD	Count
Subject Role								
Information Giver — 4 word utterances	60.7	119.1	150.5	163.9	188.2	390.9	68.63	86
Information Follower — 4 word utterances	74.0	154.5	198.3	194.2	228.6	362.0	61.39	60
Information Giver — 5 word utterances	59.2	134.9	163.4	171.3	204.9	307.1	55.74	106
Information Follower — 5 word utterances	64.6	148.7	175.2	186.2	229.4	319.5	61.74	64
Subject Gender								
Female (♀) — 4 word utterances	61.3	101.8	125.7	141.0	160.4	332.4	58.89	40
Male (♂) — 4 word utterances	60.7	146.6	187.6	189.7	226.3	390.9	65.55	106
Female (♀) — 5 word utterances	59.2	112.7	151.5	153.0	177.9	287.9	52.19	50
Male (♂) — 5 word utterances	64.6	145.7	180.8	186.8	227.3	319.5	58.09	120

Table 9: *Summary and standard deviation of 4 and 5 word utterances (wpm)*

for utterances with more words, but it is not the case for the Information Follower, Male (♂) subjects or the difference between four word Information Follower and five word Information Giver utterances. In both cases, the effect of the role of the subject (Information Giver being slower than the Information Follower) or the gender of the subject Female (♀) being slower than the Male (♂)) is stronger than the influence of the utterance word count. For this reason, the wording of "an abrupt rise of speaking rate for the segments containing from one to seven words" by Yuan et al. (2006, p. 542) would not be appropriate for the data of the ILMT-s2s corpus.

A further point of interest is the fact that Yuan et al. (2006, pp. 544) also identified the ambiguity of the wpm measurement method caused by the different "word-frequency distributions" when describing the gender speech rate difference: "It might be due to things that we would not normally think of as speech-rate parameters, such as differences in word-frequency distributions". When Figure 10

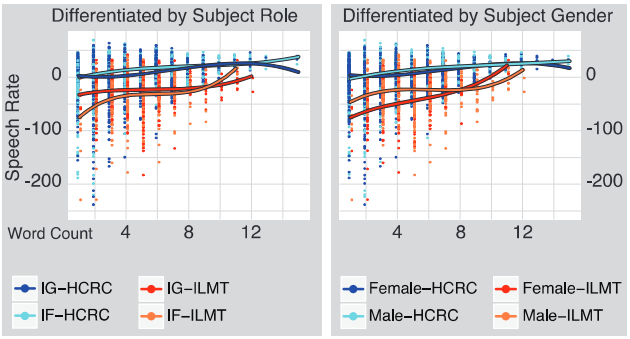

Figure 11: *Scatter plot of ILMT-s2s corpus utterance durations by word count*

which used the wpm as the speech rate indicator is represented using the speech rate calculation method of comparing the TTS output duration with the subject utterance duration, it is possible to see that the speech rate of the subjects do not show "an abrupt rise" as indicated in Figure 11 with the lines representing the HCRC Edinburgh Map Task corpus closer to a linear model than a non-linear model from 1 word all the way to 12 word utterances. This is both an indication of (i) how close the TTS output is to direct human communication, and (ii) the lack of "an abrupt rise" in the speech rate, with the word-frequency distributions no longer affecting the measurement values of the method explained in this paper.

6. Acknowledgements

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) and the Research Centres Programme (Grant 13/RC/2106) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin. The ADAPT Centre for Digital Content Technology is co-funded under the European Regional Development Fund.

7. Bibliographical References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

Arai, T. and Greenberg, S. (1997). The Temporal Properties of Spoken Japanese are Similar to those of English. In *Proceeding of EUROSPEECH '97*, pages 1011–1014, Rhodes, Greece. ISCA.

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke, V., and Niemann, H. (2001). Whence and whither prosody in automatic speech understanding: a case study. In *Proceedings of the Workshop on Prosody and Speech Recognition*, pages 3–12. ISCA.

Bell, A., Gregory, M. L., Brenier, J. M., Jurafsky, D., Ikeno, A., and Girand, C. (2002). Which Predictability Measures Affect Content Word Durations? In *Proceedings of the ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, pages 1–5. ISCA.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Boersma, P. and van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, 5(9-10):341–347.

Branigan, H. P., Pickering, M. J., Pearson, J., and McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

de Jong, N. H. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, May.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). Switchboard: telephone speech corpus for research and development. *Proceedings of ICASSP-92*, pages 517–520.

Greenberg, S. (1997). The Switchboard Transcription Project. Technical report, Johns Hopkins University, Baltimore, Maryland, USA.

Greenberg, S. (1999). Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29(2–4):159–176.

Hara, K. and Iqbal, S. T. (2015). Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study. In *Proceedings of CHI'15*, pages 3473–3482, New York, NY, USA. ACM.

Hayakawa, A., Cerrato, L., Campbell, N., and Luz, S. (2015). A Study of Prosodic Alignment in Interlingual Map-Task Dialogues. In The Scottish Consortium for ICPhS, editor, *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK. The University of Glasgow. Paper number 0760.1-9.

Hayakawa, A., Haider, F., Luz, S., Cerrato, L., and Campbell, N. (2016a). Talking to a system and oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task. In *Proceedings of Speech Prosody 2016 (SP8)*, pages 776–780, Boston, Massachusetts, USA. ISCA.

Hayakawa, A., Luz, S., Cerrato, L., and Campbell, N. (2016b). The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of LREC 2016*, pages 605–612, Paris, France. ELRA.

Hayakawa, A., Vogel, C., Luz, S., and Campbell, N. (2017a). Perception Changes With and Without the Video Channel: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. In *Proceedings of CogInfoCom 2017*, pages 401–406, Debrecen, Hungary. IEEE.

Hayakawa, A., Vogel, C., Luz, S., and Campbell, N. (2017b). Speech Rate Comparison when Talking to a System and Talking to a Human: A study from a Speech-to-Speech, Machine Translation mediated Map Task. In *Proceedings of INTERSPEECH'17*, pages 3286–3290, Stockholm, Sweden. ISCA.

Lewis, W. D. (2015). Skype translator: Breaking down language and hearing barriers. *Translating and the Computer (TC37)*.

Matsuda, S., Hu, X., Shiga, Y., Kashioka, H., Hori, C., Yasuda, K., Okuma, H., Uchiyama, M., Sumita, E., Kawai, H., and Nakamura, S. (2013). Multilingual Speech-to-Speech Translation System: VoiceTra. In *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management*, volume 2, pages 229–233, Milan, Italy. IEEE.

Morikawa, O. and Maesako, T. (1998). HyperMirror: Toward Pleasant-to-use Video Mediated Communication System. In *Proceedings of CSCW '98*, pages 149–158, New York, NY, USA. ACM.

Sztahó, D., Kiss, G., and Vicsi, K. (2015). Estimating the Severity of Parkinson's Disease from Speech Using Linear Regression and Database Partitioning. In *Proceedings of INTERSPEECH'15*, pages 498–502, Dresden, Germany. ISCA.

Yuan, J. and Jurafsky, D. (2005). Detection of questions in Chinese conversational speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 47–52, Nov.

Yuan, J., Liberman, M., and Cieri, C. (2006). Towards an Integrated Understanding of Speaking Rate in Conversation. In *Proceedings of INTERSPEECH'06*, pages 2197–2200, Pittsburgh, PA, USA. ISCA.

Zipf, G. K. (1945). The Meaning-Frequency Relationship of Words. *The Journal of General Psychology*, 33(2):251–256.

8. Language Resource References

Hayakawa, A. and Luz, S. and Cerrato, L. and Campbell, N. (2015). *The ILMT-s2s Corpus*. CNGL Programme, distributed via ELRA, Trinity College Dublin, 1.0, ISLRN 100-610-774-625-0.